Sunfall: A Collaborative Visual Analytics System for Astrophysics

Cecilia R. Aragon[1] LBNL | Stephen J. Bailey[2] LBNL | Sarah Poon[3] SSL | Karl J. Runge[4] SSL | Rollin C. Thomas[5] LBNL

Abstract

Computational and experimental sciences produce and collect ever-larger and complex datasets, often in large-scale, multi-institution projects. The inability to gain insight into complex scientific phenomena using current software tools is a bottleneck facing virtually all endeavors of science. In this paper, we introduce Sunfall, a collaborative visual analytics system developed for the Nearby Supernova Factory, an international astrophysics experiment and the largest data volume supernova search currently in operation. Sunfall utilizes novel interactive visualization and analysis techniques to facilitate deeper scientific insight into complex, noisy, high-dimensional, high-volume, time-critical data. The system combines novel image processing algorithms, statistical analysis, and machine learning with highly interactive visual interfaces to enable collaborative, user-driven scientific exploration of supernova image and spectral data. Sunfall is currently in operation at the Nearby Supernova Factory; it is the first visual analytics system in production use at a major astrophysics project.

Index Terms—Data and knowledge visualization, scientific visualization, visual analytics, visual exploration, astrophysics.

Introduction

Many of today's important scientific breakthroughs are being made by large, interdisciplinary collaborations of scientists working in geographically widely distributed locations, producing and collecting vast and complex datasets. These large-scale science projects require software tools that support, not only insight into complex data, but collaborative science discovery. Visual analytics approaches, combining statistical algorithms and advanced analysis techniques with highly interactive visual interfaces that support collaborative work, offer scientists the opportunity for in-depth understanding of massive, noisy, and high-dimensional data. Astrophysics in particular lends itself to a visual analytics approach due to the inherently visual nature of much astronomical data (including images and spectra).

One of the grand challenges in astrophysics today is the effort to comprehend the mysterious "dark energy," which accounts for three-quarters of the matter/energy budget of the universe. The existence of dark energy may well require the development of new theories of physics and cosmology. Dark energy acts to accelerate the expansion of the universe (as opposed to gravity, which acts to decelerate the expansion). Our current understanding of dark energy comes primarily from the study of supernovae [2, 3].

The Nearby Supernova Factory (SNfactory) [1] is an international astrophysics experiment designed to discover and measure Type Ia supernovae in greater number and detail than has ever been done before. These supernovae are stellar explosions that have a consistent maximum brightness, allowing them to be used as "standard candles" to measure distances to other galaxies and to trace the rate of expansion of the universe and how dark energy affects the structure of the cosmos. The SNfactory receives 50-80 GB of image data per night, which must be processed and examined by teams of domain experts within 12-24 hours to obtain maximum scientific benefit from the study of these rare and short-lived stellar events.

In order to facilitate the supernova search and data analysis process and enable scientific discovery for project astrophysicists, we developed Sunfall (SuperNova Factory AssembLy Line), a collaborative visual analytics system for the Nearby Supernova Factory that has been in production use for over a year. Sunfall incorporates sophisticated astrophysics image processing algorithms, machine learning capabilities including boosted trees and support vector machines, and astronomical data analysis with a usable, highly interactive visual interface designed to facilitate collaborative decision making. An interdisciplinary group of physicists, astronomers, and computer scientists (with specialties in machine learning, visualization, and user interface design) were involved in all aspects of Sunfall design and implementation.

Sunfall Architecture & Components

Sunfall contains four major components: Search, Workflow Status Monitor, Data Forklift, and Supernova Warehouse (Figure 1).

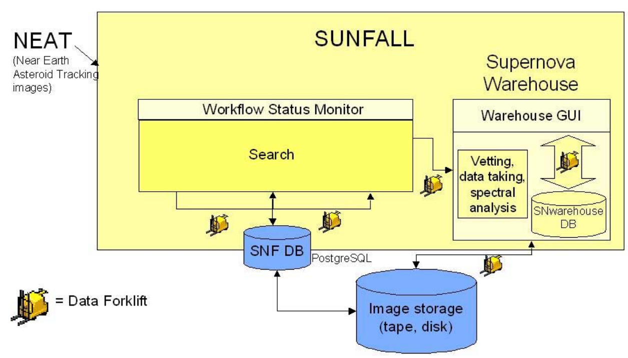

Figure 1. Sunfall architecture diagram, depicting the four components (Search, Workflow Status Monitor, Data Forklift, and Supernova Warehouse) and data flow between the components

Search incorporates SNfactory legacy software for starfield image processing and subtraction, and includes machine learning algorithms and novel Fourier contour descriptor algorithms to reduce the number of false positive supernova candidates.

Scientific workflow systems, such as the Kepler project [4], have demonstrated their usefulness in managing dataflow and visualization in complex scientific projects. The Workflow Status Monitor is a web-based monitor to facilitate collaboration and improve project scientists' situational awareness of the data flow by displaying all relevant workflow (search pipeline) status data on a single site. The supernova search software is highly parallelized as 30,000 images are queued for processing in a multi-stage pipeline that runs on the National Energy Research Scientific Computing Center (NERSC) 700-node computing cluster, the Parallel Distributed Systems Facility (PDSF). Nodes frequently go down or jobs fail, and failures must be detected promptly and jobs resubmitted quickly due to the time-critical nature of the search. Detection of such failures was challenging and time-consuming

[1]Lawrence Berkeley National Laboratory. E-Mail: aragon@hpcrd.lbl.gov
[2]Lawrence Berkeley National Laboratory. E-Mail: sjbailey@lbl.gov
[3]Space Sciences Laboratory. E-Mail: sspoon@lbl.gov
[4]Space Sciences Laboratory. E-Mail: kjrunge@lbl.gov
[5]Lawrence Berkeley National Laboratory. E-Mail: rcthomas@lbl.gov

before the deployment of the Workflow Status Monitor. The monitor displays graphs and visual displays of job completion times on PDSF, job queues, PDSF node uptimes, and disk vault loads.

The Data Forklift is a middleware mechanism consisting of a coordinator and a suite of services to automate astronomical data transfers in a secure, reliable, extensible, and fault-tolerant manner. The Data Forklift also provides the middleware for the other three components, transferring data between heterogeneous systems, databases and formats securely and reliably. The Data Forklift supports "different place, asynchronous" collaborative scientific work by facilitating data and information transfer amongst a geographically separated team. Due to the time-critical nature of the data collection (telescopes must be operated at night and are located in different time zones), tasks must take place at distinct, specified times.

The Supernova Warehouse (SNwarehouse) is a comprehensive supernova data management, workflow visualization, and collaborative scientific analysis tool. The SNwarehouse contains a PostgreSQL database, middleware consisting of Forklift mechanisms, and a graphical user interface (GUI) implemented in Java. SNwarehouse supports collaborative remote asynchronous work in several different ways.

Collaboration members can access the GUI from any networked computer worldwide via a Forklift remote deployment mechanism. Security is provided via password authentication and encrypted communication channels. SNwarehouse furnishes project scientists with a shared workspace that enables easy distribution, analysis, and access of data. Collaboration members can view, modify, and annotate supernova data, add comments, change a candidate's state, and schedule follow-up observations from work, home, while observing at the telescope, or when attending conferences. This access is critical due to the 24/7 nature of SNfactory operations. All transactions are recorded in the SNwarehouse database, and the change history and provenance of the data is permanently stored (records cannot be deleted in order to maintain the change history) and continuously visible to all authenticated users.

SNwarehouse centralizes data from multiple sources and supports task-oriented workflow. Project members perform well-defined tasks, such as vetting, scheduling, and analyzing targets, which collectively accomplish the goal of finding and following type Ia supernovae. Typically, an individual or small group performs a given task, and the results of the task provide inputs for the next task in the workflow, often performed by another set of group members. Thus, the inputs and outputs of any task must be well-defined and easily recognizable.

An example of an SNwarehouse view that supports a typical project task is given in Figure 2. Observers need to point the telescope and spectrograph at potential supernovae throughout the night, taking exposures and making judgments as to the quality of the observation. The Data Taking tool has simplified the process and reduced operator error.

This tool displays two of the custom visualizations developed for project scientists, which enable them to more easily gain insight into the accuracy of experimental data. The visualization at the top of the Data Taking display depicts three-dimensional astronomical data in an easy-to-understand two-dimensional format, with time running horizontally along the x-axis. The visualization in the lower left-hand corner enables observing scientists to detect at a glance whether the spectrograph has accurately centered the target supernova.

The raw data from the SNIFS spectrograph is complex and requires a significant amount of processing in order to yield meaning to the scientist. A visual depiction of the accuracy of the pointing and signal strength provides much more information more quickly than tables of numeric data.

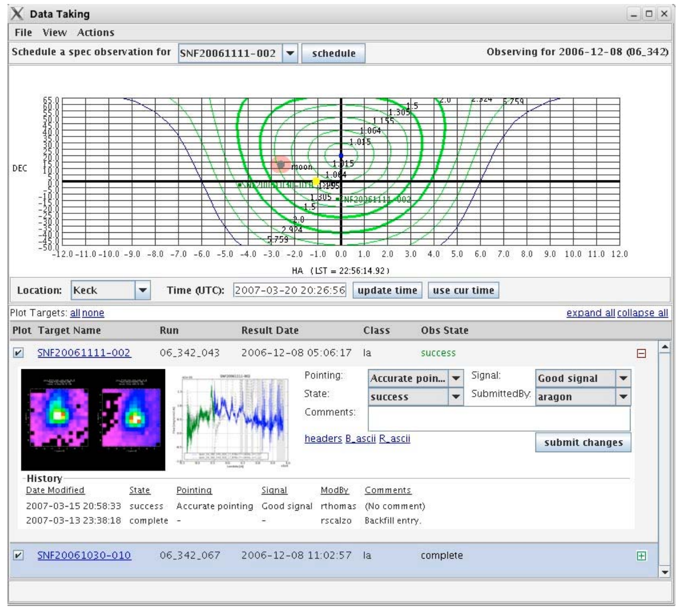

Figure 2. SNwarehouse Data Taking window. The observer can follow the targets on the Sky visualization, take notes on the success or failure of each observation, telescope status and weather conditions, and reschedule targets if necessary.

CONCLUSION

Sunfall, a collaborative visual analytics system in operation at a large-scale astrophysics project, has demonstrated that such systems that facilitate scientific analytic discourse and computer-supported collaborative work can have a positive impact on data-intensive science. In the process of design and implementation, we learned that an interdisciplinary team incorporating specialists from several fields, including scientific domain experts, will be most effective in designing an effective visual analytics system for science.

ACKNOWLEDGMENTS

We would like to thank the scientists of the SNfactory collaboration for their time and detailed feedback. This work was supported in part by the Director, Office of Science, Office of Advanced Scientific Computing Research, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098, and by the Director, Office of Science, Office of High Energy Physics, of the U.S. Department of Energy under Contract No. DE-FG02-92ER40704, and by a grant from the Gordon & Betty Moore Foundation. This research used resources of the National Energy Research Scientific Computing Center, supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

REFERENCES

[1] G. Aldering, G. Adam, P. Antilogus, P. Astier, R. Bacon, et al., "Overview of the Nearby Supernova Factory," *Proceedings of the SPIE*, vol. 4836, 2002.

[2] S. Perlmutter, G. Aldering, G. Goldhaber, et al., "Measurements of Omega and Lambda from 42 High-Redshift Supernovae," *Astrophysical Journal*, vol. 1999, pp. 565-586, 1999.

[3] A. G. Riess, A. V. Filippenko, et al., "Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant," *Astrophysical Journal*, vol. 1998, pp. 1009-1038, 1998.

[4] B. Ludaescher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger-Frank, M. Jones, E. Lee, J. Tao, and Y. Zhao, "Scientific Workflow Management and the Kepler System," Concurrency and Computation: Practice & Experience, Special Issue on Scientific Workflows, 2005.